

The chemist versus the machine: traditional knowledge versus machine learning techniques

Janine George,¹ Geoffroy Hautier^{1*}

¹ Institute of Condensed Matter and Nanosciences, Université catholique de Louvain,
Chemin des étoiles 8, 1348 Louvain-la-Neuve, Belgium

*Correspondence: geoffroy.hautier@uclouvain.be (G. Hautier)

Abstract

Chemical heuristics have been fundamental to the advancement of chemistry and materials science. These heuristics are typically established by scientists using knowledge and creativity to extract patterns from limited datasets. Machine learning offers opportunities to perfect this approach using computers and larger datasets. Here, we discuss the relationships between traditional heuristics and machine learning approaches. We show how traditional rules can be challenged by large-scale statistical assessment and how traditional concepts commonly used as features are feeding the machine learning techniques. We stress the waste involved in relearning chemical rules and the challenges in terms of data size requirements for purely data-driven approaches. Our view is that heuristic and machine learning approaches are at their best when they work together.

Keywords: Machine learning; chemical heuristics; materials discovery; materials informatics

The interplay of chemical heuristics and machine learning

Data science (see glossary), **artificial intelligence**, and **machine learning** are nowadays present in all fields of science and technology including chemistry and materials science. The impact of these techniques is expected to be very large leading to a new path towards scientific discovery (sometimes called the 4th paradigm in science) [1,2]. We will not review here all progress and future directions in the use of machine learning in materials science and we refer the interested readers to recent reviews on this topic (e.g., [3–6]). Instead, we offer a personal perspective on how these new techniques heavily relying on sophisticated algorithms and large data sets compete, complement, challenge, and/or benefit from more traditional heuristic approaches.

Learning from data: from classical heuristics to data relationships

Since the early days of chemistry, scientists have looked for patterns from often very limited data. This led to many of today's widely used chemical heuristics such as the periodic system of elements, electronegativities, and atomic radii. These heuristic models were built by combining the knowledge and creativity of the scientist with simplified physical pictures. For example, Pettifor introduced a completely new chemical scale that enabled separation of structure types of AB compounds within a 2D map. This scale was based on the size, valence, and electronegativity of the constituting atoms [7]. We will call this traditional approach classical or **chemical heuristics** throughout this article (see Figure 1, Key Figure). Recently, the traditional heuristic approach has started to be replaced by machine learning techniques. This is not only due to improvements in the machine learning methods that are now widely available through open source software but also to an ever-increasing amount of data available mainly through high-throughput *ab initio* computations [8–11]. These machine learning approaches can be divided into two categories. The first approach uses features often based on chemical heuristics (e.g., atomic radii or electronegativity) as inputs to machine learning techniques that can provide a relationship between these features and material properties. Here, the machine learning brings new relationships to properties from traditional heuristic descriptors. For instance, machine learning has been used to find a mathematical relationship

between well-known atomic features (e.g., atomic radii and ionization potential) to discriminate between wurtzite and rock salt forming binaries [12]. The second approach relies only on rules derived from relationships within the data bypassing traditional chemical descriptors. For example, the probability of certain ions to replace each other has been derived in this manner [13]. No physical/chemical feature such as atomic radius was implied in this study – only the idea that some atoms are more likely to replace other atoms than others. The three different approaches that include traditional chemical heuristics, the extraction of rules between established chemical features using machine learning, and purely data-driven approaches are summarized in Figure 1. The chemical heuristics approaches require less data but are inherently biased towards pre-conception that could turn out to be incorrect. Conversely, the purely data-driven approach could require data set sizes that are sometimes not available.

“Classical” chemical heuristics

Many of the heuristics that are still taught in general chemistry courses date back from at least a century ago. The concept of electronegativity was proposed by Avogadro and Berzelius in ~1809 [14], oxidation states by Wöhler (~1835) [15], atomic radii by Loschmidt (~1866) [16], and the periodic table of elements dates 150 years back to Mendelejew [17]. Goldschmidt and Pauling derived rules on the stability of crystal structures more than 90 years ago [18,19]. Without any doubts, chemical heuristics have been instrumental to advances in chemistry and materials sciences over the last hundred years. However, one must be careful to not use these heuristics blindly. Their historical importance should not exclude them from critical assessment. Rahm and colleagues have recently shown that two of our most fundamental chemical heuristics – the periodic table and electronegativities – can change drastically at high pressures [20]. Here, we face the common issue of extrapolation. For example, Li changes its valence and becomes a p group element at 300 GPa. K and heavier alkali metals become transitional metals. Furthermore, Na becomes the most electropositive s^1 element. Thus, the structure of the periodic table and chemical reactivities will change completely for some elements with rising pressure. Similar changes should be expected for typical oxidation states. Chemical heuristics might have only limited transferability, needing to be adapted to allow for the exploration of more extreme

conditions. Even in more common conditions, very well-established rules can turn out to be less powerful than expected when evaluated with modern techniques and larger data sets. Hautier and colleagues have recently evaluated the predictive power of the Pauling rules which connects the coordination environments of a crystal with its stability. For instance, the 1st Pauling rule links the preferred coordination environment of a cation to the cation-anion ratio (see Figure 2a). Our analysis on oxides shows that Pauling's first rule is only fulfilled for 66% of all tested local environments with important deviations in alkali and alkali-earth chemistries for instance (Figure 2a). Strikingly, and despite being a corner stone of solid-state chemistry, only 13% of a data set of 5000 oxides fulfil the four other Pauling rules [21]. This new assessment could only be performed using modern tools of information technology. First, a diverse set of oxide crystal structures had not only to be determined by crystallographers but also made readily available in databases such as the Inorganic Crystal Structure Database (ICSD), the Open Crystallographic Database, or the Cambridge Structural Database [22–24]. Without these efforts of the crystallographic community, many of today's data-driven searches would not be possible. Furthermore, automatic tools for the determination of coordination environments had to be developed [25,26], and a quantitative assessment of the Pauling rules had to be set-up [21]. In today's world of unprecedented access to information and automation, it is staggering to realize that only several hundred crystal structures had been determined when Pauling proposed his rules and that no computer was available to process the data [27].

In a similar spirit of challenging established rules, Filip and Giustino have shown that a modified and improved Goldschmidt tolerance factor that relates the ionic radii of the ions building a perovskite structure to the stability of perovskites has predictive power (Figure 2b) [28]. The Goldschmidt tolerance factor is based on the idea of a tightly packed structure of ionic spheres. To improve the basic idea of the tolerance factor, Filip and Giustino included further geometric constraints in their prediction that were initially missing in the definition of Goldschmidt, based on the idea of a high stability of tightly packed structures. This improved tolerance factor can be used to predict stable perovskites with a high fidelity (80%). Almost needless to say that the dataset of stable perovskite structure available to Filip and Giustino was much larger than the one to Goldschmidt in 1926. Based on this improved tolerance factor, Filip and Giustino predicted 90.000 hitherto unknown perovskites.

Since the initial development of most chemical rules, both the amount of data available and computing processing power have grown substantially. Now is the time for the community to test chemical heuristics more widely and rigorously. This includes assessing their validity in less common conditions (e.g., high pressure and/or temperature) and developing new rules following the spirit of previous heuristics but updated using modern machine learning techniques and larger data sets.

Physically or chemically inspired description of materials in machine learning

Going further along the data arrow from Figure 1, another popular approach has been to use concepts from the classical heuristics as features or descriptors for machine-learning studies of materials properties (Figure 3) [3]. The material is here represented by traditional physical and chemical features such as electronegativities or atomic radii, and machine learning models that link these descriptors to materials properties are trained on large data sets. To get back to the previous examples of the Pauling rules, such approaches could, for instance, start from similar descriptors (e.g., local environments or atomic radii) but search through machine learning for relationships that deviate from these empirical rules. Various materials properties such as the band gap, bulk and shear modulus, and vibrational properties have successfully been machine learned in this way [3]. These studies can link composition to properties including, if necessary, crystal structure features. For instance, researchers have shown that vibrational entropy could be inferred from composition and crystal structure features including traditional heuristic descriptors such as atomic numbers, radii, position in the periodic table, Pettifor scales, or electronegativities (Figure 4a) [29,30]. While materials descriptions based solely on composition were the norm in the earlier days, there is a growing set of methods that very elegantly incorporate structural information. For example, graph-based structure representations have shown very promising results [31,32]. For instance, crystal graph convolutional neural networks allowed learning of different properties (e.g., formation energy or band gaps) for a variety of structure types and compositions [31]. These **crystal graph convolutional neural networks** are based on a representation of atoms as nodes and bonds as edges within a graph. Additional information on the atom and bonds that are partially based on chemical heuristics (e.g., position in periodic table or electronegativities) are included in node and edge feature vectors. On top of this graph, a convolutional neural

network is built that allows to automatically extract the optimal representations of target properties. In a similar way, Chen and colleagues introduced MatErials Graph Networks (MEGNet) to represent crystals and molecules within machine-learning models of material properties [32].

There is an abundance of chemical features to use and efforts have started to group them in easily accessible codes and databases (matminer [33]). However, not all descriptors are adequate to predict a property. While previous chemical knowledge could help choosing descriptors (e.g., the Pauling rules could motivate to focus on atomic radii and local environments to predict material stability), it is also possible to find the adequate descriptors from data only. The approach is here to start with many descriptors or features and select the most relevant ones. This is a very common task implemented in many machine learning methods [12,34,35]. Work from Ghiringhelli and colleagues [12] illustrates the importance of selecting the right descriptors by performing a classification of AB binaries between zinc blende and rock salt formers. Using the **least absolute shrinkage and selection operator (LASSO)** which helps selecting descriptors in a large pool of typical atomic features, the authors discovered what descriptors would be the most predictive and establish a simple relationship to be used for classification (Figure 4b). Here, they found that atomic ionization potentials, electron affinities and radii describing where the radial probability density of the valence s or p orbitals show its maximum were the most effective descriptors on a large pool of features. Figure 4b also illustrates that the relationship between the selected descriptor and the target property is relatively simple. Feature selection can also be efficiently combined with less-interpretable feed-forward neural networks on crystal structure graphs [36]. The importance of feature selection for these neural networks has been shown to be critical when data sets are small (around hundreds of data points).

Solely learning from data

The third approach of machine learning is to directly extract relationships between the raw data without using the intermediate step of chemical or atomistic features. For instance, one would like to directly link the crystal structure description (position and chemical elements on the sites) to properties. This strategy is appealing as it does not

require knowledge *a priori* of what features or descriptors would be predictive, but it usually requires much more data (Figure 1).

An interesting example of such a data-driven approach comes from the field of crystal structure prediction. Based on atomic radii and groups in the periodic table, chemists have foreseen which ions can substitute each other for more than 100 years. In recent years, a model based on machine learning that can quantify how likely certain cations can substitute to each other has been developed [13]. Essentially, the model builds a probability function providing the likelihood for two ions or elements to substitute to each other. No assumption is made on what chemical features (ionic radius, electronegativity, position in the periodic table) drives these substitutions. The likelihood of substitution between two ions is given as a matrix in Figure 5. This algorithm was successfully used and combined with density functional theory (DFT) to predict new Li-ion battery, luminescent, or ternary nitride materials [37–39] that were confirmed experimentally. Some of those substitutions were surprising and not easily rationalizable on standard chemical grounds. For instance, the substitution model suggested a Ba^{2+} with Sr^{2+} , Re^{7+} with Al^{3+} , and N^{3-} with O^{2-} substitution to form a new phosphor host material $\text{Sr}_2\text{LiAlO}_4$ which was successfully synthesized [37].

Another field where data-driven models are replacing traditional physical models is interatomic potentials [5,40,41]. Traditional force-fields such as Lennard-Jones potentials [42] or embedded atom models (EAM) [43] have strong physical assumptions embedded. Recently developed ML potentials bypass these assumptions by directly connecting energetics and forces to crystal structure without much functional assumption about their relationships. In a sense, they are purely data-driven with minimal physical assumptions. To fit these potentials, a large set of reference data (e.g. from DFT) is needed. There are several types of these ML interatomic potentials: they differ by the descriptor used to encode the atomic potentials and the type of regression. Popular ones are the Gaussian Approximation Potentials by Bartók and Csányi where a “Smooth Overlap of Atomic Positions” (SOAP) descriptor is used and the Neuronal Networks by Behler where atom centred symmetry functions are used [40,41,44,45]. The SOAP descriptor encodes atomic geometries with the help of a local expansion of gaussian smeared atomic densities. The possibilities offered by these potentials are immense allowing simulations of

phenomena requiring many atoms (e.g., grain boundaries [46,47] or amorphous phases [48,49]) or accelerating the search of new crystal structures [50]. It is expected that they will impact many fields and have already been used in the investigation of battery and potential thermoelectric materials [51,52].

While these data-driven approaches are attractive because they do not rely on descriptors that could turn out to be inadequate, we should keep in mind that they could simply re-learn well-established chemical rules. An illustrative example comes from a search for water-splitting materials among the ABX_3 perovskites [53]. Using a genetic algorithm, a search among all perovskites was performed focusing on stability but also band gap, and band edges. This search was purely data-driven using no heuristics (Figure 6, blue line (best GA)). The authors performed another search on the same data set including constraints from typical chemical rules: the oxidation states in one material should sum up to 0, only materials with an even number of electrons in the primitive unit cell are considered to avoid metals, and the classical Goldschmidt tolerance factor is used for ranking of the materials. The search based on traditional chemical heuristics was similarly successful as the genetic algorithms (Figure 6, red line (chemical)). Obviously here, the machine had been “reinventing the wheel” or more precisely re-learning known rules such as charge balance. Interestingly, it is the combination of data-driven optimization and heuristics that lead to the best performances. Joining the two approaches together prevents to reinvent the wheel while capturing relationships that might be unknown from traditional knowledge (Figure 6 green line). One of the earlier mentioned graph-based approaches (MEGNet) allows to combine the purely data-driven approach with heuristic features as well and it enables the researcher to decide on the amount of traditional heuristic knowledge (from only atomic numbers and simple bonding information to the inclusion of more complex chemical heuristics) used in the developed models depending on the data availability [32].

Concluding Remarks

Chemical heuristics and machine learning are rooted in the same desire to build models explaining patterns in data. Instead of seeing traditional chemical heuristics and machine learning studies only as competitors, we think they instead feed on each

other. Machine learning approaches can benefit from old chemical heuristics as they provide natural features to represent materials (see Outstanding Questions). The traditional chemical rules also provide standard benchmarks for machine learning studies. Researchers should be encouraged to not only validate their machine learning models by cross-validation but also compare the results and performance of artificial intelligence to traditional rules. In our opinion, machine learning provides a true added value only when it convincingly beats traditional approaches. If the outcome of a new neural network is to simply re-learn the rule of charge balance, one can question the advance in knowledge achieved by the machine learning tool. Chemical heuristics should also be challenged by machine learning and large-scale statistical studies. The historical significance of a heuristic rule should not preclude its critical assessment. In fact, researchers should remember that many traditional concepts and rules in chemistry have been developed on relatively small datasets and might, therefore, be highly biased. The boom of information technology presents an opportunity for chemists to get back to the drawing board and design new rules based on large datasets with the help of machine learning approaches.

Acknowledgements

J. G. acknowledges funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 837910. We thank Pierre-Paul De Breuck for helpful comments on the manuscript. Figure 4 was created with the help of VESTA [57].

Conflict of interest

The authors declare no conflict of interest.

Glossary

Artificial Intelligence: Artificial intelligence refers to “intelligent” behaviour of machines. Machine learning can be considered a subfield of artificial intelligence.

Chemical Heuristics: Chemical heuristics are intuitive strategies that guide chemists in solving problems and making decisions [55].

Crystal Graph Convolutional Neural Networks: Crystal Graph Convolutional Neural Networks consist of a crystal graph, where nodes represent the atoms (and their properties)

and the edges represent the connections between the atoms, and a convolutional neural network built on top of the crystal graph [31]. The latter consists of convolutional and pooling layers which produce the overall feature vector of the crystal [31] by convoluting the feature vectors of neighbouring atoms.

Data science: Data science focuses on algorithms and principles to extract patterns from large amounts of data [54]. It is closely related to the fields of machine learning and data mining and these terms are sometimes used interchangeably [54].

Goldschmidt Tolerance Factor: The Goldschmidt tolerance factor indicates the stability of crystals with perovskite structure types based on the ionic radii of ions constituting the crystal structure [19, 28].

LASSO: Least absolute shrinkage and selection operator (LASSO) is a regression with a penalty in the form of a L_1 -norm which ensures both selection and shrinkage of the variables in the regression [56].

Machine learning: Machine learning provides algorithms that can analyse large data sets and it is therefore a very important subfield of data science [54].

Pauling Rules: The Pauling rules are a structure-stability relationship and they connect the stability/existence of crystals to the local environments of cations and the way those local environments are connected [18].

References

- [1] Agrawal, A. Choudhary, A. (2016), Perspective: Materials Informatics and Big Data: Realization of the “Fourth Paradigm” of Science in Materials Science. *APL Materials*, 4, 053208.
- [2] Hey, T., Tansley, S., Tolle, K. (2009), The Fourth Paradigm: Data-Intensive Scientific Discovery; *Microsoft Research*.
- [3] Schmidt, J., Marques, M. R. G., Botti, S., Marques, M. A. L. (2019), Recent Advances and Applications of Machine Learning in Solid-State Materials Science. *npj Comput Mater.*, 5, 1–36.
- [4] Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., Walsh, A. (2018), Machine Learning for Molecular and Materials Science. *Nature*, 559 (7715), 547–555.
- [5] Deringer, V. L., Caro, M. A., Csányi, G. (2019), Machine Learning Interatomic Potentials as Emerging Tools for Materials Science. *Adv. Mater.* 31, 1902765.

- [6] Schleder, G. R., Padilha, A. C. M., Acosta, C. M., Costa, M., Fazzio, A. (2019) From DFT to Machine Learning: Recent Approaches to Materials Science—a Review. *J. Phys. Mater.* 2, 032001.
- [7] Pettifor, D. G. (1984), A Chemical Scale for Crystal-Structure Maps. *Solid State Communications*, 51, 31–34.
- [8] Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., Persson, K. A. (2013), Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.*, 1, 011002.
- [9] Draxl, C., Scheffler, M., (2019), The NOMAD Laboratory: From Data Sharing to Artificial Intelligence. *J. Phys. Mater.*, 2, 036001.
- [10] Curtarolo, S., Setyawan, W., Wang, S., Xue, J., Yang, K., Taylor, R. H., Nelson, L. J., Hart, G. L. W., Sanvito, S., Buongiorno-Nardelli, M., Mingo, N., Levy, O., (2012), AFLOWLIB.ORG: A Distributed Materials Properties Repository from High-Throughput Ab Initio Calculations. *Comput. Mater. Sci.*, 58, 227–235.
- [11] Álvarez-Moreno, M., de Graaf, C., López, N., Maseras, F., Poblet, J. M., Bo, C. (2015), Managing the Computational Chemistry Big Data Problem: The loChem-BD Platform. *J. Chem. Inf. Model.*, 55, 95–103.
- [12] Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C., Scheffler, M. (2015), Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.*, 114, 105503.
- [13] Hautier, G., Fischer, C., Ehlacher, V., Jain, A., Ceder, G. (2011), Data Mined Ionic Substitutions for the Discovery of New Compounds., *Inorg. Chem.*, 50, 656-663.
- [14] Jensen, W. B., (1996), Electronegativity from Avogadro to Pauling: Part 1: Origins of the Electronegativity Concept. *J. Chem. Educ.*, 73, 11.
- [15] Karen, P. (2015), Oxidation State, A Long-Standing Issue! *Angew. Chem. Int. Ed.*, 54, 4716–4726.
- [16] Rahm, M., Hoffmann, R., Ashcroft, N. W. (2016), Atomic and Ionic Radii of Elements 1 -96. *Chem. Eur. J.*, 22, 14625–14632.
- [17] Mendelejew, D. (1869), Über die Beziehungen der Eigenschaften zu den Atomgewichten der Elemente, *Zeitschrift für Chemie*, 12, 405–406.
- [18] Pauling, L. (1929), The Principles determining the structure of complex ionic crystals, *J. Am. Chem. Soc.*, 51, 1010.

- [19] Goldschmidt, V. M. (1926), Die Gesetze der Krystallochemie. *Naturwissenschaften*, 14, 477-485.
- [20] Rahm, M., Cammi, R., Ashcroft, N. W., Hoffmann, R. (2019), Squeezing All Elements in the Periodic Table: Electron Configuration and Electronegativity of the Atoms under Compression. *J. Am. Chem. Soc.*, 141, 10253–10271.
- [21] George, J., Waroquiers, D., Di Stefano, D., Petretto, G., Rignanese, G., Hautier, G., (2020), The Limited Predictive Power of the Pauling Rules. *Angew. Chem. Int. Ed.*, 59, 7569–7575.
- [22] Groom, C. R., Bruno, I. J., Lightfoot, M. P., Ward, S. C. (2016), The Cambridge Structural Database. *Acta Cryst B*, 72, 171–179.
- [23] Zagorac, D., Müller, H., Ruehl, S., Zagorac, J., Rehme, S. (2019), Recent Developments in the Inorganic Crystal Structure Database: Theoretical Crystal Structure Data and Related Features. *J Appl Cryst*, 52, 918–925.
- [24] Gražulis, S., Chateigner, D., Downs, R. T., Yokochi, A. F. T., Quirós, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P., Le Bail, A. (2009) Crystallography Open Database – an Open-Access Collection of Crystal Structures. *J. Appl. Cryst.*, 42, 726–729.
- [25] Waroquiers, D., Gonze, X., Rignanese, G.-M., Welker-Nieuwoudt, C., Rosowski, F., Göbel, M., Schenk, S., Degelmann, P., André, R., Glaum, R., Hautier, G. (2017), Statistical Analysis of Coordination Environments in Oxides. *Chem. Mat.* 29, 8346.
- [26] Waroquiers, D., George, J., Horton, M., Schenk, S., Persson, K. A., Rignanese, G.-M., Gonze, X., Hautier, G. (2020), ChemEnv: A Fast and Robust Coordination Environment Identification Tool. *Acta Cryst B*, 76, 683–695.
- [27] Behrens, H. (1996), Data Import and Validation in the Inorganic Crystal Structure Database. *J. Res. Natl. Inst. Stand. Technol.* 101, 365.
- [28] Filip, M. R., Giustino, F., (2018), The Geometric Blueprint of Perovskites. *PNAS*, 115, 5397–5402.
- [29] Legrain, F., Carrete, J., van Roekeghem, A., Curtarolo, S., Mingo, N., (2017), How Chemical Composition Alone Can Predict Vibrational Free Energies and Entropies of Solids. *Chem. Mater.*, 29, 6220–6227.
- [30] Tawfik, S. A., Isayev, O., Spencer, M. J. S., Winkler, D. A. (2020), Predicting Thermal Properties of Crystals Using Machine Learning. *Adv. Theor. Simul.* 3, 1900208.

- [31] Xie, T., Grossman, J. C. (2018), Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.*, 120, 145301.
- [32] Chen, C., Ye, W., Zuo, Y., Zheng, C., Ong, S. P. (2019), Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.*, 31, 3564–3572.
- [33] Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N. E., Bajaj, S., Wang, Q., Montoya, J., Chen, J., Bystrom, K., Dylla, M. (2018), Matminer: An Open Source Toolkit for Materials Data Mining. *Comput. Mater. Sci.* 152, 60–69.
- [34] Ghiringhelli, L. M., Vybiral, J., Ahmetcik, E., Ouyang, R., Levchenko, S. V., Draxl, C., Scheffler, M. (2017), Learning Physical Descriptors for Materials Science by Compressed Sensing. *New J. Phys.*, 19, 023017.
- [35] Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M., Ghiringhelli, L. M. (2018), SISO: A Compressed-Sensing Method for Identifying the Best Low-Dimensional Descriptor in an Immensity of Offered Candidates. *Phys. Rev. Materials*, 2, 083802.
- [36] De Breuck, P.-P., Hautier, G., Rignanese, G.-M. (2020) Machine Learning Materials Properties for Small Datasets. arXiv:2004.14766 [cond-mat].
- [37] Wang, Z., Ha, J., Kim, Y. H., Im, W. B., McKittrick, J., Ong, S. P. (2018) Mining Unexplored Chemistries for Phosphors for High-Color-Quality White-Light-Emitting Diodes. *Joule*, 2, 914–926.
- [38] Chen, H., Hautier, G., Jain, A., Moore, C., Kang, B., Doe, R., Wu, L., Zhu, Y., Tang, Y., Ceder, G. (2012), Carbonophosphates: A New Family of Cathode Materials for Li-Ion Batteries Identified Computationally. *Chem. Mater.*, 24 (11), 2009–2016.
- [39] Sun, W., Bartel, C. J., Arca, E., Bauers, S. R., Matthews, B., Orvañanos, B., Chen, B.-R., Toney, M. F., Schelhas, L. T., Tumas, W., Tate, J., Zakutayev, A., Lany, S., Holder, A. M., Ceder, G. (2019), A Map of the Inorganic Ternary Metal Nitrides. *Nat. Mater.*, 18, 732–739.
- [40] Bartók, A. P., Payne, M. C., Kondor, R., Csányi, G. (2010), Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.*, 104, 136403.
- [41] Behler, J., Parrinello, M. (2007), Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.*, 98, 146401.

- [42] Jones, J. E., Chapman, S. (1924), On the Determination of Molecular Fields. —
II. From the Equation of State of a Gas. *Proc. Math. Phys. Eng. Sci.*, 106, 463–
477.
- [43] Daw, M. S., Foiles, S. M., Baskes, M. I. (1993), The Embedded-Atom Method: A
Review of Theory and Applications. *Materials Science Reports*, 9 (7), 251–310.
- [44] Bartók, A. P., Kondor, R., Csányi, G., (2013), On Representing Chemical
Environments. *Phys. Rev. B*, 87, 184115.
- [45] Behler, J. (2011), Atom-Centered Symmetry Functions for Constructing High-
Dimensional Neural Network Potentials. *J. Chem. Phys.* 134, 074106.
- [46] Hu, C., Zuo, Y., Chen, C., Ping Ong, S., Luo, J. (2020), Genetic Algorithm-
Guided Deep Learning of Grain Boundary Diagrams: Addressing the Challenge
of Five Degrees of Freedom. *Materials Today*, 38, 49-57.
- [47] Yokoi, T., Noda, Y., Nakamura, A., Matsunaga, K. (2020), Neural-Network
Interatomic Potential for Grain Boundary Structures and Their Energetics in
Silicon. *Phys. Rev. Materials*, 4, 014605.
- [48] Sosso, G. C., Miceli, G., Caravati, S., Behler, J., Bernasconi, M. (2012), Neural
network interatomic potential for the phase change material GeTe, *Phys. Rev.*
B, 85, 174103.
- [49] Deringer, V. L., Csányi, G., (2017), Machine Learning Based Interatomic
Potential for Amorphous Carbon. *Phys. Rev. B*, 95, 094203.
- [50] Deringer, V. L., Proserpio, D. M., Csányi, G., Pickard, C. J. (2018), Data-Driven
Learning and Prediction of Inorganic Crystal Structures. *Faraday Discuss.*, 211,
45–59.
- [51] Deringer, V. L. (2020), Modelling and Understanding Battery Materials with
Machine-Learning-Driven Atomistic Simulations. *J. Phys. Energy*, DOI
10.1088/2515-7655/abb011.
- [52] George, J., Hautier, G., Bartók, A. P., Csányi, G., Deringer, V. L. (2020),
Combining Phonon Accuracy with High Transferability in Gaussian
Approximation Potential Models. *J. Chem. Phys.*, 153, 044104.
- [53] Jain, A., Castelli, I. E., Hautier, G., Bailey, D. H., Jacobsen, K. W. (2013),
Performance of Genetic Algorithms in Search for Water Splitting Perovskites. *J*
Mater Sci, 48, 6519.
- [54] Kelleher, J. D., Tierney, B. (2018), Data Science, *MIT Press*.

- [55] Graulich, N., Hopf, H., Schreiner, P. R. (2010), Heuristic Thinking Makes a Chemist Smart. *Chem. Soc. Rev.*, 39, 1503–1512.
- [56] Tibshirani, R. (2011), Regression shrinkage and selection via the lasso: a retrospective. *J R Stat Soc Series B*, 73, 273-282.
- [57] Momma, K., Izumi, F. (2011), VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data, *J. Appl. Crystallogr.*, 44, 1272-1276.

Figure Captions.

Figure 1. A hierarchy inside of models based on data and chemical insight of the model.

Figure 2. Assessment and modification of classical chemical heuristics. a) Assessment of the first Pauling rule, adapted with permission from [21]. (Left) Illustration of the rule. (Right) Percentages of correct predictions by Pauling's rule to all tested predictions. b) Illustration of the updated Goldschmidt tolerance factor by Filip and Giustino, reprinted with permission from [28]. Copyright 2008 National Academy of Sciences. (Left) Illustration of the different geometric stability limits for the constituting ions. (Right) Graphical depiction of the newly derived stability area of ternary perovskites (blue) by imposing the no-rattling principle and other limits (chemical limits CL1 and CL2, octahedral limit OL, stretch limit SL, secondary stretch limits SSL1 and SSL2, and tilt limits TL1 and TL2)

Figure 3. Figure exemplifying the representation of the composition and/or structure of a material to simplify the learning process of materials properties.

Figure 4. Importance of the descriptor for machine learning. a) The figure shows the mean absolute error for the prediction of different vibrational properties. 5 different types of descriptors are tested: they are based on the chemical composition, elemental properties of the atoms, pair correlation functions (PCF), solid angles as defined by O'Keefe, and bispectrum components based on the local atomic density around each atom. Reprinted with permission from [29]. Copyright 2017 American Chemical Society. The figure is licensed under a CC BY 4.0 license. b) Energy differences between rock salt and zinc blend structures of binary AB materials as a function of the optimal two-dimensional descriptor that was identified in [12]. Here, the descriptor clearly separates the compounds preferring rock salt and zinc blend structure type within a 2D plot. Reprinted with permission from [12]. The figure is licensed under a CC BY 3.0 license.

Figure 5. Likelihood of ions to substitute each other. Here, a pair correlation function for each ion couple is shown. The tendency to substitute each other is indicated by a positive value, the tendency to not substitute each other is indicated by a negative value. Reprinted with permission from [13]. Copyright 2011 American Chemical Society.

Figure 6. Comparison of potential solar-water splitting materials derived by a random search (random), a search based on well-known chemical heuristics (chemical), a search by a genetic algorithm (GA), and a combination of the chemical heuristic with the genetic algorithm (best GA + chemical). Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature J. Mater. Sci. [53], Copyright 2013.

Figures.

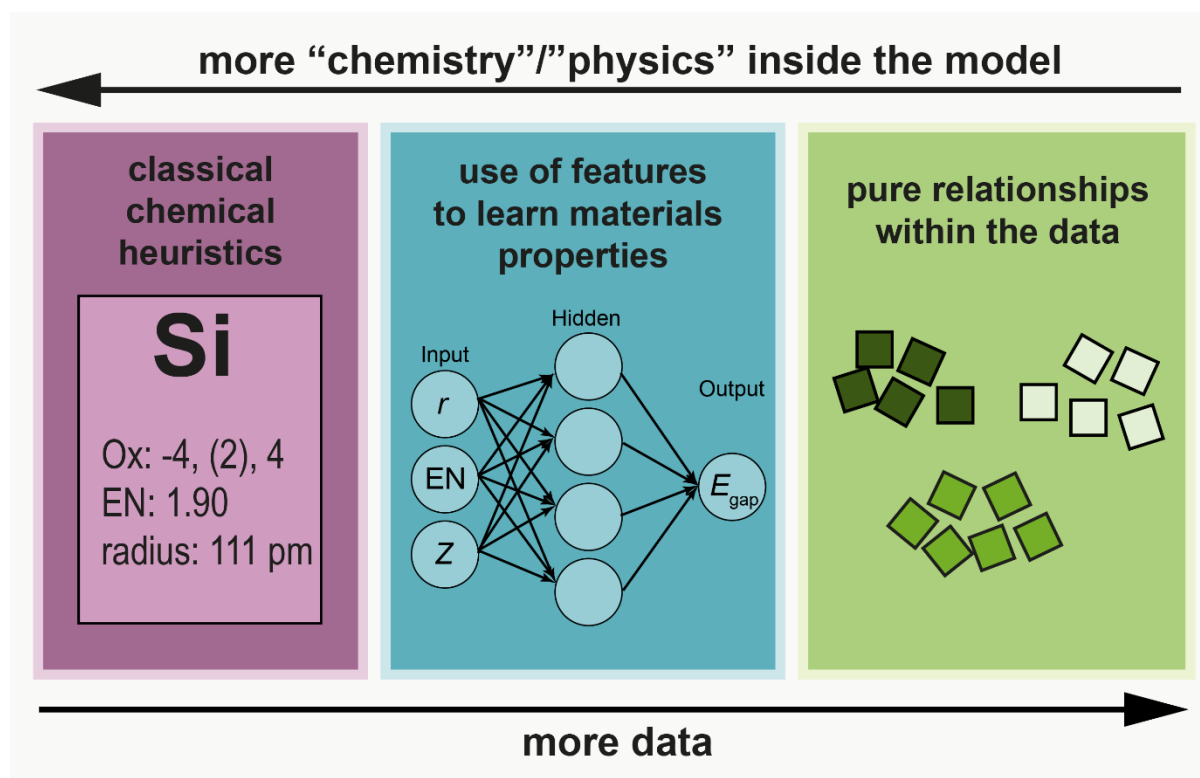
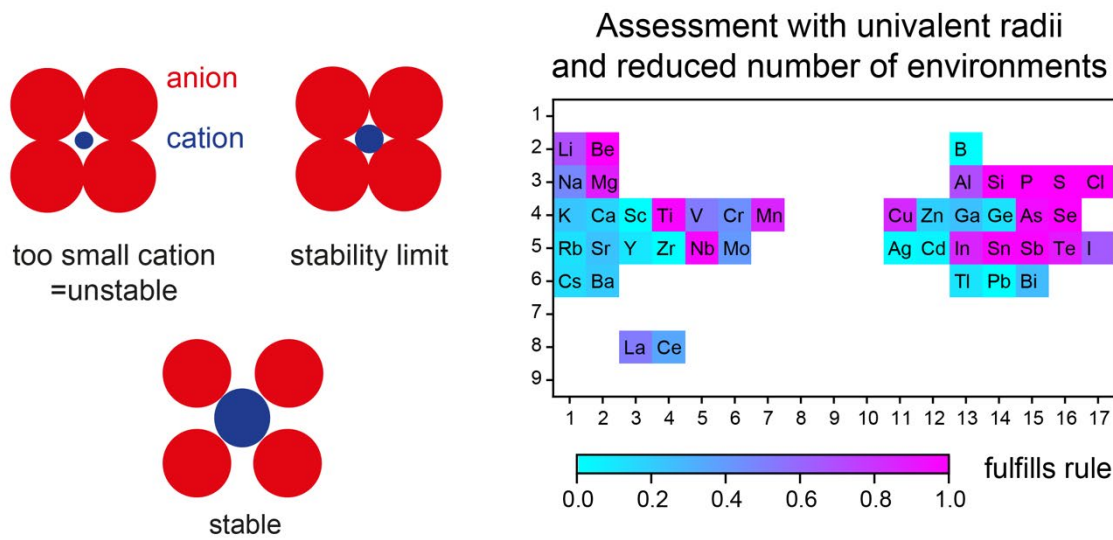
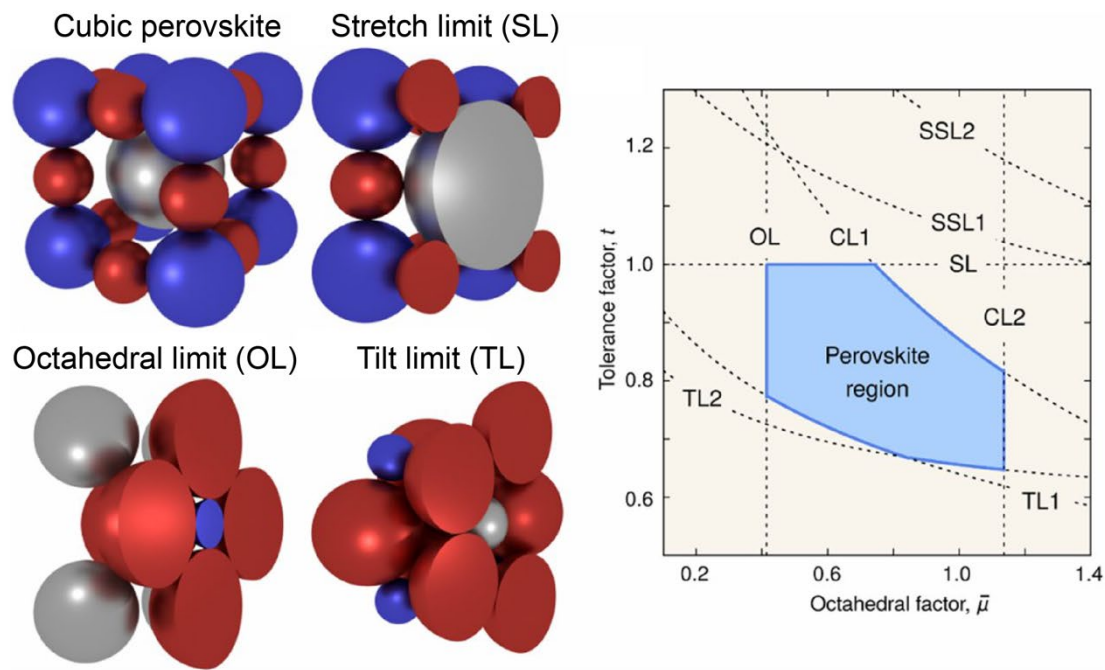


Figure 1.

a) Pauling rules: Assessment of the first rule



b) Goldschmidt tolerance factor



526

527 **Figure 2.**

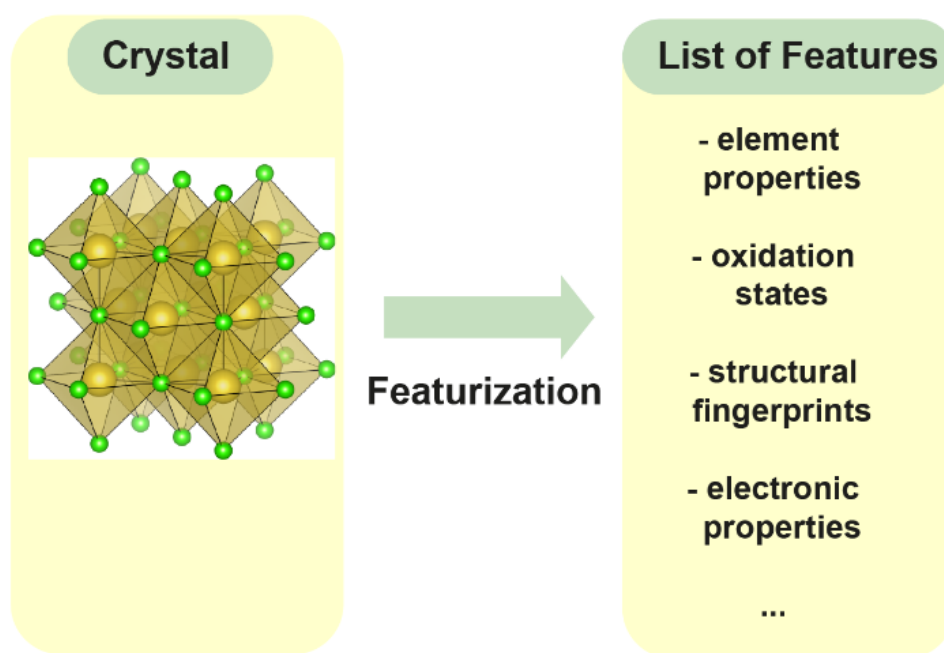
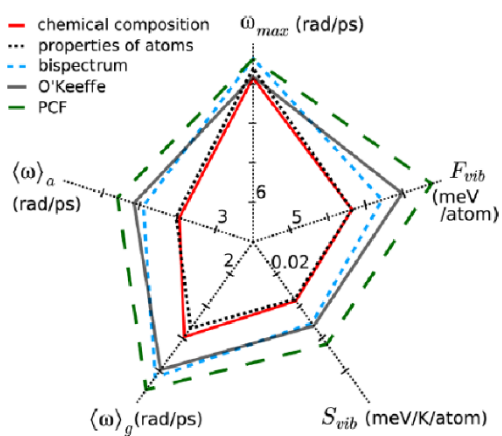


Figure 3.

a)



b)

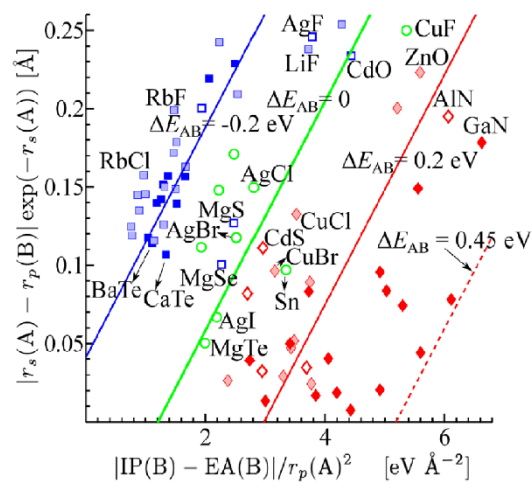


Figure 4.

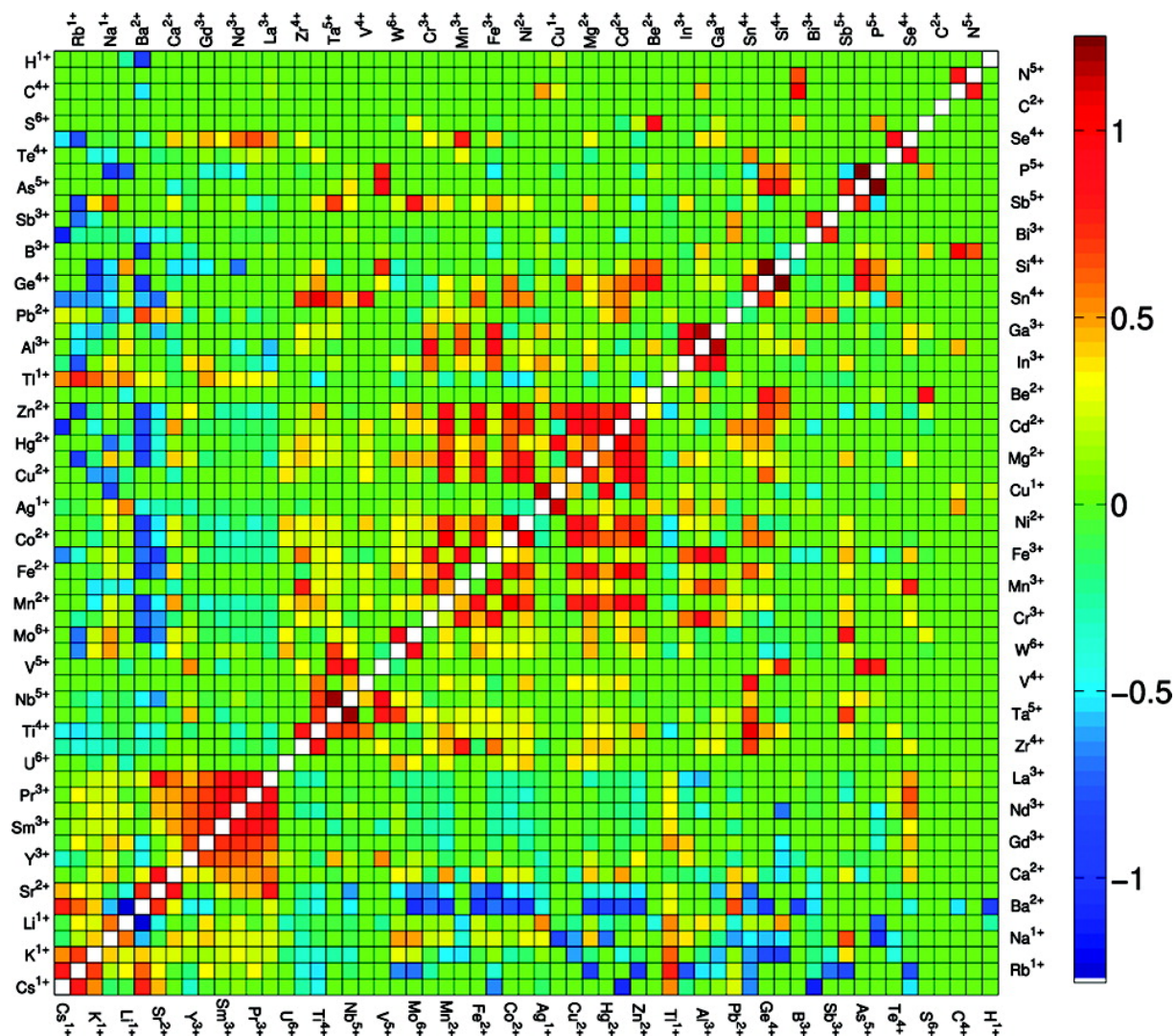


Figure 5.

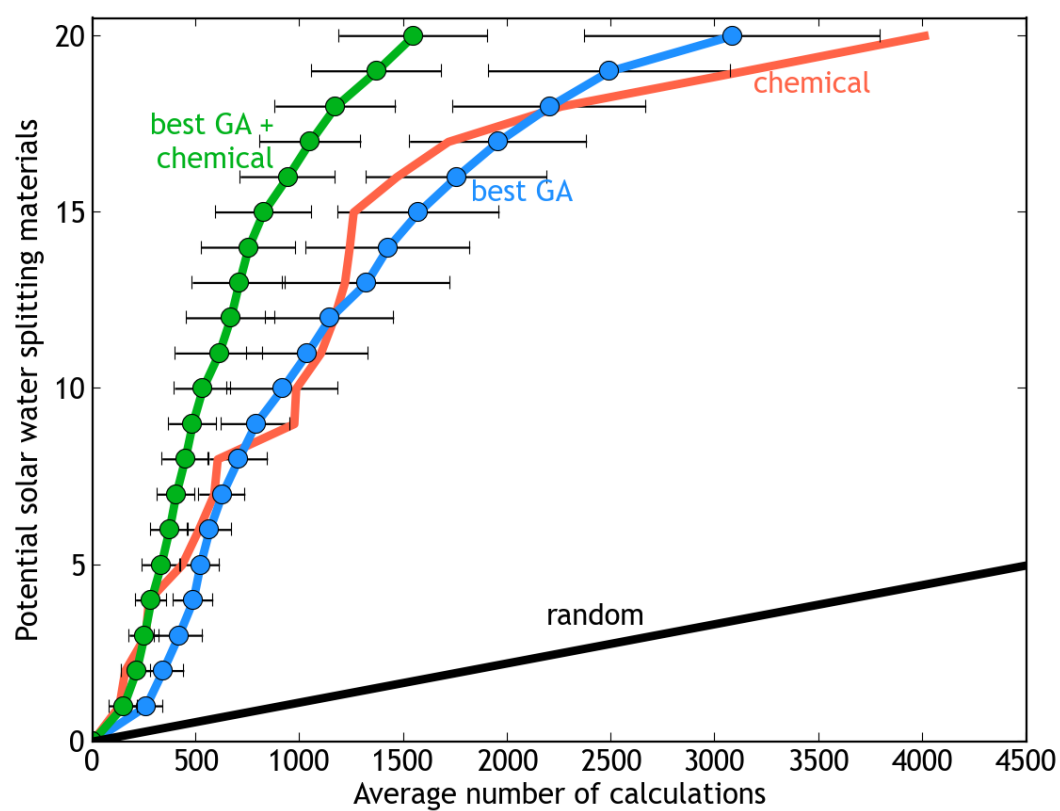


Figure 6.